Universal Segmentation at Arbitrary Granularity with Language Instruction

Yong Liu[1] , Cairong Zhang[2] , Yitong Wang[2] , Jiahao Wang[3] , Yujiu Yang[1] Yansong Tang[1]*
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]ByteDance Inc. [3]The University of Hong Kong
liuyong23@mails.tsinghua.edu.cn, tang.yansong@sz.tsinghua.edu.cn

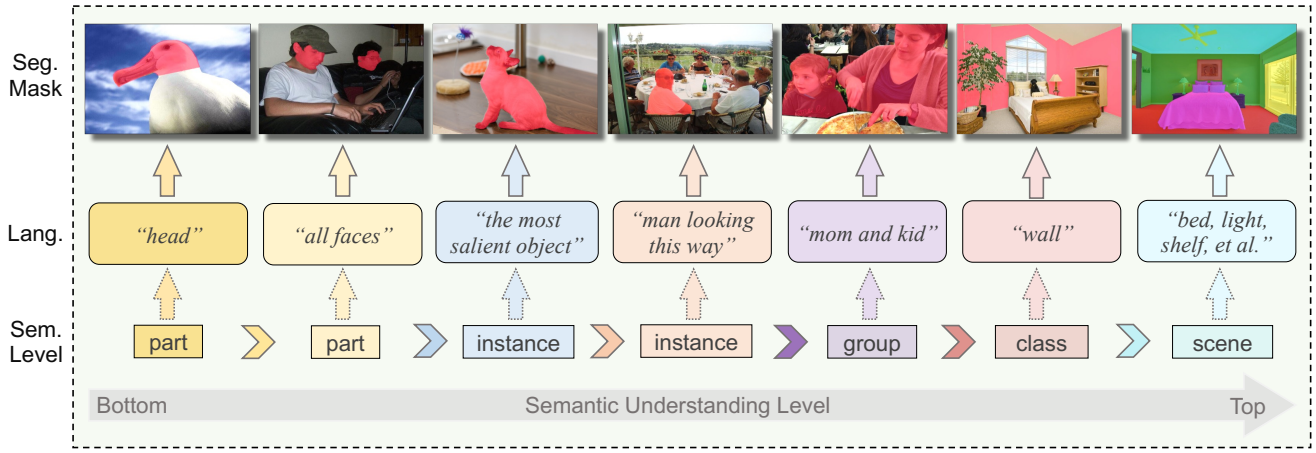


Figure 1. Illustration of our UniLSeg that is able to segment images at any granularity or semantic level with language as instructions. "Seg. Mask", "Lang.", and "Sem. Level" denote the segmentation masks, corresponding language descriptions, and semantic levels, respectively. The segmentation masks are shown in red or other colors. UniLSeg can take arbitrary text as input, whether it is a detailed long description of an object or a short category name. With flexible expressions indicating segmentation target, UniLSeg achieves excellent performance on various semantic level, *e.g.*, object part, single or multiple instances, and the whole scene.

Abstract

This paper aims to achieve universal segmentation of arbitrary semantic level. Despite significant progress in recent years, specialist segmentation approaches are limited to specific tasks and data distribution. Retraining a new model for adaptation to new scenarios or settings takes expensive computation and time cost, which raises the demand for versatile and universal segmentation model that can cater to various granularity. Although some attempts have been made for unifying different segmentation tasks or generalization to various scenarios, limitations in the definition of paradigms and input-output spaces make it difficult for them to achieve accurate understanding of content at arbitrary granularity. To this end, we present UniLSeg, a universal segmentation model that can perform segmentation at any semantic level with the guidance of language instructions. For training UniLSeg, we reorganize a group of tasks from original diverse distributions into a unified data format, where images with texts describing segmentation targets as input and corresponding masks are output. Combined with a automatic annotation engine for utilizing numerous unlabeled data, UniLSeg achieves excellent performance on various tasks and settings, surpassing both specialist and unified segmentation models. Code is available here.

1. Introduction

Segmentation is one of the most important problem in computer vision, which aims to group meaningful regions and perform pixel-level understanding. Recent years have witnessed great progress in the development of various segmentation tasks such as semantic segmentation [7, 16, 46, 48], interactive segmentation [30, 44, 45, 74], salient object segmentation [26, 85], and referring segmentation [51, 78].

Although many excellent works have emerged, they tend to be specialist approaches for specific segmentation tasks, making it difficult for them to address complex and diverse segmentation scenarios. When adapting to novel settings or semantics, new models need to be designed and trained

[1]Corresponding author

on data of corresponding distribution, which leads to significant data and computation cost. Therefore, it is greatly promising to achieve versatile and universal segmentation that can cater to various semantic levels and settings. Nevertheless, due to the diverse distribution of data and the complexities of input-output space, designing and training such a model presents a significant challenge.

Recently, some works [30, 67, 75, 86] have attempted to propose unified paradigm for multiple segmentation tasks or generalization to various scenarios. Among them, SAM [30] and SEEM [86] propose to take point as the basis for indicating segmentation targets and have achieved impressive performance and generalization ability. However, such point-based interaction paradigm tends to produce over-dispersed and unexpected segmentation results due to the lack of semantic concept awareness. Besides, the information contained within the "point" is insufficient to guide the model in executing generic segmentation across multiple semantic levels. Unlike SAM, UNINEXT [75] focuses on object-centric segmentation and adopt prompt generation to standardize the input space. Despite achieving excellent results, the emphasis on instance makes it difficult to achieve understanding of any granularity, such as fine-grained part segmentation and coarse-grained scene understanding. With visual in-context learning, Painter [66] and SegGPT [67] realize flexible target-aware segmentation. But the unification of input-output example form of different tasks has brought up new challenges and greatly limits the application scenarios.

Therefore, it leads to a natural question: *is there a unified paradigm that can conveniently interact with the model for universal segmentation at any granularity and has good scalability?* The answer we proffer is language. As the most important tool for humanity, language can flexibly express the objects of reality and the laws of thought. Like communication between human beings, language has become one of the most important means of communication between humans and machines. It has the capability to provide information at various levels of detail, effectively guiding the model in accomplishing the desired tasks. Such versatile and informative prompt is in line with the above requirements for unified segmentation instruction. With this insight, we study a series of tasks to validate our ideas. The explored tasks contain referring image segmentation (RIS) [68, 78], semantic segmentation (SS) [7, 48], salient object detection (SOD) [56, 57], part segmentation (PS) [17, 32], referring video object segmentation (RVOS) [51, 60], and open-vocabulary segmentation (OVS) [16, 73]. We reorganize these tasks from original diverse distributions into a unified data format, where images with texts describing segmentation targets as input and corresponding masks are output. Benefiting from such flexible and unified design, segmentation model can be jointly trained on different tasks to learn the connections between language instructions and visual concepts.

In addition, to promote the the model's understanding of high-level language instructions, we present a fully aligned framework called UniLSeg. The core design philosophy of UniLSeg lies in conducting extensive visual-linguistic interaction to identify and segment the targets within multimodal joint space. As shown in Figure 1, with generic representations learned from numerous data of various tasks, our UniLSeg is able to comprehend diverse language expressions that indicate segmentation targets at varying semantic levels. Furthermore, to exploit the large-scale unlabelled and weakly supervised data, we propose an automatic annotation engine to generate pseudo caption-mask pairs for assisted training. Extensive experiments on a group of benchmarks demonstrate the powerful segmentation ability of our UniLSeg, *e.g.*, it overpasses specialist models and unified competitors by about 12% and 7% on G-Ref [54] validation set.

Our contributions can be summarized as follows:

- We present UniLSeg, a generic segmentation model that fully integrates visual concept with textual guidance. With language instructions as universal prompt, UniLSeg can be jointly trained on multiple tasks to learn the connections between diverse textual descriptions and visual content, achieving universal segmentation at arbitrary semantic granularity.
- Our UniLSeg achieves superior performance on a group of challenging benchmarks from various tasks, which benefits from the language-based paradigm definition and additional automatic annotation engine for large-scale unlabeled and weakly-labeled data.

2. Related Work

Unified Large Segmentation Model Some recent works have been devoted to exploring unified and robust large segmentation models. Among them, SAM [30] has received great attention for its powerful and generalizable segmentation ability. SAM proposes to leverage point-based prompts for indicating target regions. A series of follow-up works [21, 32, 76, 81] have built on SAM, applying it to a variety of tasks and achieving outstanding performance. However, random point sampling and point-based interaction may lead to over-dispersed segmentation masks and unawareness of high-level semantics. In addition, Painter [66] and SegGPT [67] leverage visual in-context learning and reorganize the output space of different tasks for unified prediction. UNINEXT [75] and SEEM [86] design unified segmentation decoder for adapting different task prompts. Although these works have achieved excellent performance, limitations in the definition of paradigms and input-output space make it difficult for them to perform segmentation at any semantic granularity.

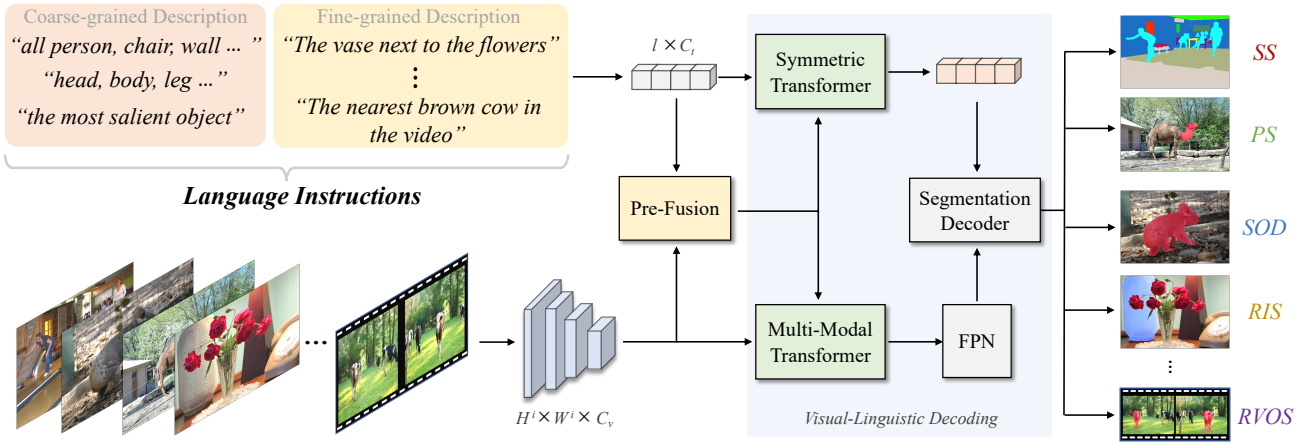


Figure 2. Pipeline of our UniLSeg. It takes both images and corresponding language prompt as input. With versatile language descriptions indicating segmentation targets and full visual-linguistic interactions, UniLSeg can perform segmentation at any semantic granularity and tackle various tasks such as semantic segmentation (SS), part segmentation (PS), salient object detection (SOD), open-vocabulary segmentation (OVS), referring image (RIS) and video object segmentation (RVOS).

Language-Guided Segmentation As the most representative language-guided segmentation task, referring image segmentation aims to perform pixel-level visual-linguistic alignment for input images and given descriptions. The pioneer works [18, 39] extracts image and language features respectively and concatenates them to form the multi-modal features. The subsequent approaches generally can be divided into two categories. The first idea [20, 77, 80] is utilizing the internal structure of the text to help identify target objects. However, this approach does not model well-aligned cross-modal joint space, and the pipeline tends to be complex. The other idea [4, 9, 13, 19, 24, 27, 63, 68, 78, 79] is to model the cross-modal relations between image and language by various attention operations. Following the attention alignment idea, there are many follow-up works such as LAVT [78], GRES [40], and PolyFormer [41]. In addition to referring segmentation, semantic segmentation can also be taken as language-guided task and the category names can be viewed as short and rough text descriptions [15, 16, 31, 73]. Different from the detailed descriptions in referring segmentation, prior guidance provided by class name is coarse-grained, which poses challenges in multi-modal interaction. Even with the help of additional vision-language models such as CLIP [58], the model performance is still unsatisfactory.

3. Method

3.1. Pipeline

Overview: Figure 2 shows the pipeline of our UniLSeg. The core design philosophy of UniLSeg lies in performing extensive visual-linguistic interaction, which is compatible with the language-based unified paradigm. Specifically, it takes both images and corresponding language prompt as input. By perceiving segmentation target in cross-modal joint space and activating corresponding response, UniLSeg achieves universal segmentation of arbitrary semantic granularity, represented by a group of tasks. We will elaborate it in the following sections.

Encoding Process: For the input image $I \in \mathbb{R}^{H \times W \times 3}$, we utilize a pyramidal vision encoder [47] to extract the hierarchical vision feature $f_v^i \in \mathbb{R}^{H^i \times W^i \times C_v^i}$, $i \in$ [1,2,3,4]. Here H^i and W^i denote the height and width of i-th scale feature map, respectively. C_v denotes the channel dimension of visual features.

For the input language prompt $L \in \mathbb{R}^l$, we take the transformer-based language encoder [58] to encode it to a word embedding $f_w \in \mathbb{R}^{l \times C_t}$ and an overall sentence embedding $f_s \in \mathbb{R}^{1 \times C_t}$, where l is the length of the input language expression. The word embedding f_w contains fine-grained guidance information. The sentence embedding f_s, on the other hand, expresses the general characteristics of segmentation targets. Joint utilization of f_w and f_s contributes to different tasks.

Pre-Fusion: Pre-Fusion aims to incorporate language guidance into visual features and roughly highlight target areas, which helps to mitigate the impact of background noise to visual-linguistic joint space. During our exploration, we have discovered that this module does not require a complex design. Implemented with simple multi-head cross-attention, this part can achieve desired activation effect. Specifically, Pre-Fusion takes the word feature f_w and hierarchical vision feature f_v^i, $i \in$ [2,3,4] as input. Here take i-th scale as an example for illustration. Since the purpose of this structure is to elicit potential response within visual

content, we take the visual feature f_v^i as the Query and the word embedding f_w as the Key and Value. The process can be formulated as:

$$f_c^i = softmax(\frac{G_q(f_v^i)^T G_k(f_w)}{\sqrt{C^i}})G_v(f_w)^T, \quad (1)$$

where G_q, G_k, G_v are projection functions that transfer the input features to the corresponding space. C denotes channel dimension of joint embedding space. f_c is activated visual feature used for subsequent visual-linguistic decoding.

Visual-Linguistic Decoding: We design a two-stream decoding structure to fully utilize the guidance from language instructions and align cross-modal joint space. Specifically, in the vision path, the visual features are sent to a Multi-Modal Transformer for learning intra- and inter-modality connections. In the language path, we incorporate the activated visual context into linguistic prompt embedding to generate content-aware prompts. This approach enhances the alignment between visual and linguistic spaces and effectively reduces cross-modal domain gaps. Below we present the details of these two paths.

Vision Path: This path consists of a Multi-Modal Transformer and a FPN [38]. Due to the limited alignment capacity of Pre-Fusion, the multi-modal transformer is used to perform sufficient intra- and inter-modality interactions. It is mainly composed of multi-modal self-attention and multi-modal cross-attention operation. Taking the $i\text{-}th$ level visual feature f_c^i as an example, we first flatten it along the spatial dimension and add fixed positional embeddings [3] to it. After that, the flattened visual tokens are concatenated together with word embeddings f_w to form multi-modal tokens f_m. Then multi-head self-attention is applied to extract relevant information between them. Self-attention allows the model to excavate information within respective modalities while modeling the visual-linguistic joint space. Note that only the output vision tokens are used for subsequent process, and the word tokens are discarded. This process can be formulated as:

$$f_m^i = flatten(f_c^i) + Pos., \quad (2)$$

$$f_m^i = Concat(f_m^i, f_w), \quad (3)$$

$$f_b^i = MHSA(f_m^i)[: H^i W^i], \quad (4)$$

$$f_b^i = LN(f_b^i) + f_c^i, \quad (5)$$

where $MHSA$ and LN denote multi-head self-attention and layer normalization [1] operation, respectively. Pos is the fixed sinusoidal positional embeddings. After that, we further leverage the output vision tokens f_b^i as Query and the word embedding f_w as Key and Value for multi-head cross-attention, benefiting the location of target regions. Finally, a FPN-like [38] structure is utilized to integrate the aligned visual features of all scales.

Table 1. Task-specific language prompt designs.

Task	Prompt Template
Referring Image Segmentation	natural caption
Referring Video Object Segmentation	natural caption
Salient Object Detection	"the most salient object"
Semantic Segmentation	"all {}"
Open-Vocabulary Segmentation	"all {}"
Part Segmentation	"all {}"

Language Path: Inspired by prompt learning [14, 59, 84], we also utilize a language instruction updating strategy to adjust linguistic space with visual content. For simplicity and elegance of structure, we still take attention operation to achieve this goal. Actually, the process of language path is symmetric with the multi-modal transformer in vision path, *i.e.*, symmetric transformer. It first takes cross-attention with sentence-level textual embedding f_s as Query and the activated visual features f_c as Key and Value. After that, self-attention is used to fully integrate initial language prompt with the content-aware one.

Finally, the activated visual features and content-aware linguistic embedding are combined to generate response map by similarity calculation, *i.e.*, matrix multiplication. With bi-linear interpolation and binarization, the model generates the output mask.

3.2. Task-Specific Prompt Design

As the most flexible prompt, language descriptions can be reorganized to fit different goals. For tasks discussed in this paper, we design specific prompt templates for them and the summary is shown in Table 1.

Referring image segmentation and referring video object segmentation aim to segment objects from images and videos based on a given language description. Since the related expressions already exist in these task, we directly leverage them as the language prompt. Semantic segmentation and open-vocabulary segmentation can be reformulated as language-guided paradigm by replacing output layers with computing the similarity between visual and linguistic embeddings. An intuitive approach is to take the category names as short textual expression. However, we find that this may potentially create semantic conflicts with long texts from referring segmentation and affects the effectiveness of joint training. To better combine data from different distributions, we design the language prompt of such category-based tasks to "all {}", where {} denotes the target category name. For salient object detection, we directly utilize "the most salient object" as the input template.

3.3. Automatic Annotation for Unlabeled Data

To train our UniLSeg and make it capable of universal segmentation of arbitrary semantic granularity, we collect and reorganize an amount of supervised training data

from available benchmarks. Specifically, these "supervised data" come from RefCOCO [25], GRefCOCO [40], COCO-Stuff [37], Ref-YouTubeVOS [60], PartImageNet [17], LIP [35], ECSSD [64], and DUTS [65]. We convert them to unified format, *i.e.*, the triplet of image, mask, and corresponding language caption. Captions for data of different tasks are defined according to the template described above.

In addition to these supervised data, we also try to leverage the numerous unlabeled images and weakly annotated data. In particular, we classify these weakly supervised or unlabeled data into three categories: box-labeled, mask-labeled, and unlabeled. Since we take the language as unified prompt, we design an automatic annotation engine to generate desired caption-mask pairs and filter out potential noise. 1) For box-labeled data that is mainly collected from Object365 [62] and RefCOCO [25], we crop related sub-images with annotated bounding boxes and then leverage SAM [30] and BLIP [33] to generate pseudo masks and captions. 2) For mask-based data sampled from SA-1B dataset [30], a naive method is captioning each annotated mask region with image-caption models. However, we find that this would lead to category redundancy and severe mismatches between different masks of the same class. To address this problem, we abandon the approach that labeling text based on the existing masks, and instead re-label both the masks and text from scratch. Specifically, we first leverage RAM [81] to tag any common categories in images. With all categories exist in image as input, Grounding DINO [43] is introduced to detect object regions related to each class tag. Then, with bounding boxes generated by Grounding DINO as box prompt, we use the huge version of SAM [30] to generate fine-grained masks for all potential categories. 3) For unlabeled data, *e.g.*, images from ImageNet [8], we first use BLIP [33] to generate natural captions for each image. Then a referring segmentation model pre-trained on referring datasets [25] is applied to recognize and segment the object related to the caption, obtaining the desired mask-caption pairs.

To remove undesired low-quality triplets and reduce the impact of annotation noise exist in pseudo-labeled data, we first leverage CLIP [58] to calculate the matching score of each caption with corresponding mask region. By filtering triplets with excessively low matching score, the quality of these pseudo-labeled data is improved. Besides, we incorporate the hide-and-seek strategy [61] during the training process to effectively alleviate the detrimental effects of inaccurate pseudo labels. The patch hiding probability is 0.2.

4. Experiment

4.1. Implementation Details.

We take the text encoder of CLIP ViT-B/16 [58] and Swin Transformer [47] pre-trained on ImageNet [8] as our encoder in default. Both language and vision encoder are initialized using the official pre-trained weights. The rest of weights in our model are randomly initialized. The input images are resized to 480 × 480 by default and no data augmentation technique is applied. We adopt a two-stage pre-training strategy to fully utilize data of different distributions. For the first stage, we train the model on 192 Tesla V100 GPUs with the batch size of 8 for each GPU. Training data of the first stage is sampled from SA-1B dataset [30]. The learning rate is set to 5e-5 and the epoch is set to 5. For the second stage, the training data consist of supervised data collected from a group of benchmarks and remain pseudo labeled data. We train the model for 15 epochs in this stage with learning rate of 1e-4. The learning rate is decayed by 0.1 after 10-th epoch. Besides, we also set a smaller learning rate for the visual backbone and the scaling factor is 0.1. The model is optimized with the combination of cross-entropy loss and Dice loss [52] with the Adam [28] optimizer. For convenience, we randomly sample one expression for each object within one iteration. During inference, the output mask is upsampled to the size of the input image by bi-linear interpolation. We binarize the prediction masks by the threshold of 0.5 and do not utilize other post-process operations.

4.2. Main Results

We evaluate our UniLSeg on 6 tasks with corresponding evaluation metrics. The inference benchmarks contain RefCOCO [25], RefCOCO+ [25], G-Ref [54], Ref-YouTubeVOS [60], ADE20K [83], Pascal Context [12], PartImageNet [17], ECSSD [64], SOD [57], and Pascal-S [53]. Note that we report the performance of two versions of UniLSeg, *i.e.*, UniLSeg-20 and UniLSeg-100, which differ in sampling proportion of the SA-1B dataset during pre-training. The results are presented below.

Referring Image Segmentation: Referring image segmentation task is the most appropriate measure for language-based segmentation because of the diversity and flexibility of the captions in this task. We compare our method with both proprietary approaches [40, 41, 68, 78] in RIS field and large unified segmentation models [75, 86]. The results are shown in Table 2. It can be seen that UniLSeg surpasses all existing methods by a significant margin, *e.g.*, 79.27 *vs* 73.41 on G-Ref [54] validation set and 73.18 *vs* 70.04 on RefCOCO+ [25] validation set.

Referring Video Object Segmentation: Results shown in Table 3 demonstrate the effectiveness of our UniLSeg in language-guided video segmentation. UniLSeg handles videos in a frame-by-frame manner. Actually, such performance stems in part from the lack of consideration of inter-frame relationships in the current RVOS benchmarks. However, robust image-level understanding ability also indicates

Table 2. Comparison with state-of-the-art methods in terms of oIoU on the popular referring image segmentation benchmarks RefCOCO [25], RefCOCO+ [25], and G-Ref [54]. We compare our UniLSeg with both proprietary approaches of RIS field and strong large segmentation models on both validation and test split. "-" represents that the result is not provided.

Method	Vision Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
Proprietary Methods									
LSCM [22]	ResNet101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
CMPC+ [42]	ResNet101	62.47	65.08	60.82	50.25	54.04	43.47	-	-
MCN [50]	DarkNet53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
EFN [13]	ResNet101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
BUSNet [77]	ResNet101	63.27	66.41	61.39	51.76	56.87	44.13	-	-
CGAN [49]	DarkNet53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
LTS [24]	DarkNet53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [9]	DarkNet53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
ResTR [27]	ViT-B	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-
CRIS [68]	ResNet50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS [68]	ResNet101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
LAVT [78]	Swin-B	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
GRES [40]	Swin-B	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
PolyFormer [41]	Swin-B	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
Unified Segmentation Models									
SEEM [86]	FocalT	-	-	-	-	-	-	65.60	-
UniNext [75]	ConvNext-L	80.32	82.61	77.76	70.04	74.91	62.57	73.41	73.68
UniLSeg-20	Swin-B	80.52	81.83	78.43	72.70	77.02	66.99	78.41	79.47
UniLSeg-100	Swin-B	**81.74**	**83.17**	**79.85**	**73.18**	**78.29**	**68.15**	**79.27**	**80.54**

Table 3. Results of referring video object segmentation on Ref-YouTubeVOS validation set.

Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [60]	47.2	45.3	49.2
LBDT-4 [10]	49.4	48.2	50.6
MTTR [2]	55.3	54.0	56.6
VLT [9]	63.8	61.9	65.6
ReferFormer [70]	59.4	58.0	60.8
SOC [51]	62.4	61.1	63.7
UniLSeg-20	64.1	61.9	66.3
UniLSeg-100	**64.9**	**62.8**	**67.0**

Table 4. Results of open-vocabulary semantic segmentation with mIoU as metric. LMM denotes large multi-modal model.

Method	Additional LMM	ADE20K-150	PC-59
LSeg+ [31]	×	13.0	36.0
SimSeg [73]	✓	20.5	47.7
OpenSeg [15]	✓	21.1	42.1
GKC [16]	×	18.8	45.2
MaskCLIP [11]	✓	23.7	45.9
OVSeg [34]	✓	24.8	53.3
UniLSeg-20	×	27.6	54.3
UniLSeg-100	×	**29.5**	**56.7**

the potential for extending our UniLSeg to corresponding video-level model.

Semantic Segmentation: Expressions in referring segmentation tend to be long texts that describes an object in detail. To prove UniLSeg has the ability to understand diverse linguistic descriptions and perform high-level scene understanding, we test it on semantic segmentation task under both out-vocabulary and in-vocabulary settings with the coarse-grained category name as language instructions.

Open-Vocabulary Setting: we directly take the pre-trained UniLSeg for open-vocabulary inference on ADE20K-150 [83] and Pascal Context-59 [12] datasets. From Table 4 we can see that our UniLSeg achieves excellent performance for open-vocabulary settings. It is worth noting that UniLSeg even outperforms the methods utilizing additional large multi-modal models such as ALIGN [23] and CLIP [58] for region classification.

In-Vocabulary Setting: we finetune the UniLSeg on ADE20K [83] dataset to evaluate it for the in-vocabulary setting. The results are shown in Table 5. Although our approach performs worse than the specialist models due to its uncustomized design about this task, it surpasses previous unified model that utilizes visual information for guidance, *e.g.*, SegGPT [67], by a remarkable margin.

Salient Object Detection: With the intuitive prompt "the most salient object", our UniLSeg achieves the best performance on all popular salient object detection benchmarks. Table 6 shows the corresponding comparison with F_{mean}

Table 5. Results of semantic segmentation with mIoU as metric.

Method	ADE20K
FCN+ [48]	29.4
DeepLabV3+ [5]	44.1
RefineNet [36]	40.7
SegFormer [72]	51.1
MaskCLIP [6]	51.1
SegGPT [67]	39.9
UniLSeg-20	45.2
UniLSeg-100	49.5

Table 6. Comparisons with salient object detection methods.

Method	ECSSD	SOD	PASCAL-S
F3Net [69]	0.912	0.775	0.816
MINET [55]	0.911	-	0.809
GateNet [82]	0.894	-	0.797
ICON [85]	0.936	0.802	0.854
MFABA [71]	0.935	-	0.857
RCSBNet [26]	0.927	-	0.842
UniLSeg-20	0.954	0.857	0.881
UniLSeg-100	**0.961**	**0.863**	**0.889**

Table 7. Results on part segmentation benchmark.

Method	Val IoU	Test IoU
SemanticFPN [29]	56.76	54.57
DeepLabV3+ [5]	60.57	58.71
SegFormer [67]	61.97	61.46
UniLSeg-20	62.46	62.03
UniLSeg-100	**63.87**	**63.62**

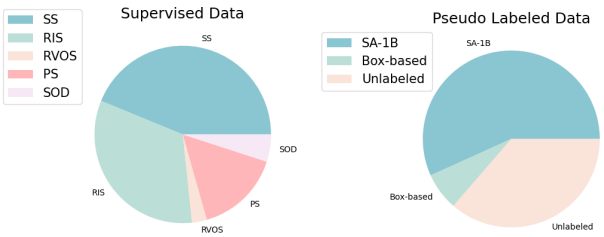
as metric. The larger is better.

Part Segmentation: The tasks above are basically instance-level or category-level. To prove the effectiveness of our method on segmenting images at any spatial granularity, we evaluate UniLSeg on the large-scale part segmentation benchmark PartImageNet [17] and the results are shown in Table 7.

4.3. Training Source Component

In this part we analyze the composition of training data. Figure 3 (a) shows the proportions of supervised data collected from different tasks. From the linguistic perspective, language expressions from RIS and RVOS are natural long linguistic captions. That from other tasks are coarse-grained short expressions, *e.g.*, category names. From the visual standpoint, RIS, RVOS, and SOD are instance-level understanding. SS and PS are about semantic-level (scene-level) and local part-level, respectively. The total number of supervised images and mask-caption pairs is 360k and 7.58M.

Figure 3 (b) demonstrates the component of pseudo labeled training source generated from weakly annotated and unlabeled data by the automatic annotation engine. "SA-1B" is corresponding to the mask-based data. The figure present the statistic for 20% SA-1B. Under such circumstances, we totally collect 3.5M images with 22M pseudo mask-caption pairs. For 100% SA-1B data, the total number is about 11.5M images with 126M mask-caption pairs.



(a) (b)

Figure 3. Illustration of training data component. (a) shows the proportions of supervised source collected from different tasks. (b) demonstrates the component of pseudo labeled training source.

Table 8. Ablation studies about the model components. The experiments are performed on RIS, SS, and SOD tasks with RefCOCO, ADE20k and SOD benchmarks, respectively.

Method	RefCOCO	ADE20K	SOD
Baseline	48.68	21.89	0.671
+PF	65.38	37.45	0.784
+PF, +LP	74.47	43.12	0.811
+PF, +LP, +VP	**77.56**	**44.63**	**0.827**

4.4. Ablation Study

Architecture Analysis To verify that the model design of our UniLSeg is compatible with the language prompt paradigm, we perform ablation experiments on different parts of the network and the results are shown in Table 8. PF, LP, and VP indicate Pre-Fusion, Vision Path, and Language Path, respectively. We take RIS, SS, and SOD tasks for evaluation. It is evident that the model's performance on the pertinent tasks enhances as the synergy between language and visual content intensifies.

Training Source Analysis This part shows the impact of different training source on segmentation performance.

Effectiveness of SA-1B data: SA-1B [30] contains a large number of high quality images from diverse scenarios. However, due to the lack of textual labels, we need to leverage existing models to generate pseudo mask-caption pairs, which leads to annotation noise and attenuates its effect. We have tried two strategies, *i.e.*, joint training and pre-training, for incorporating it into the training process. Figure 5 shows the effect of sampling 20% and 100% SA-1B data into training process under these two strategies.

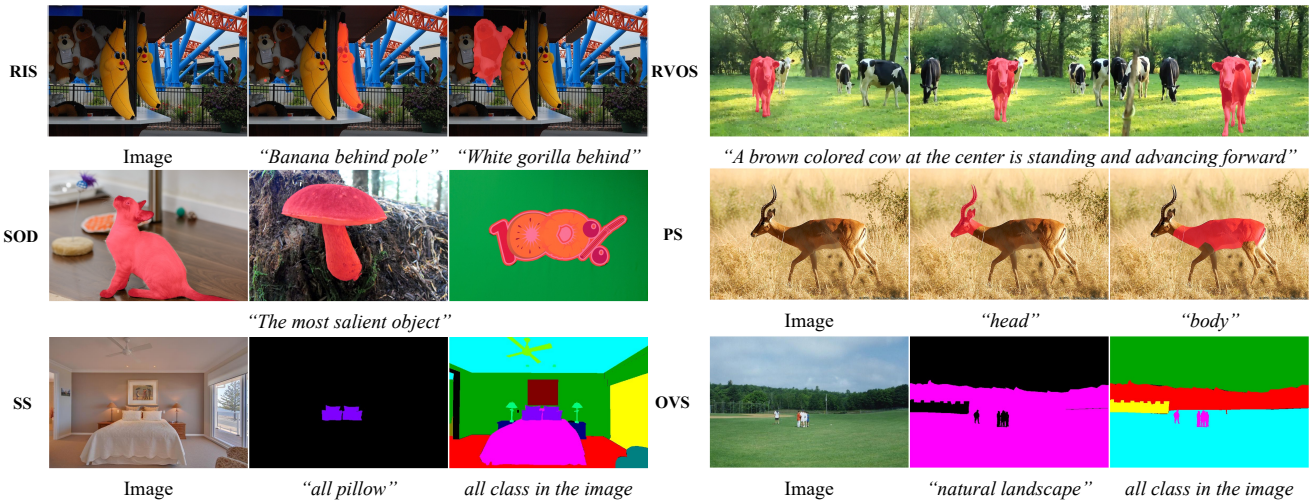


Figure 4. Visualization of segmentation results for different tasks.

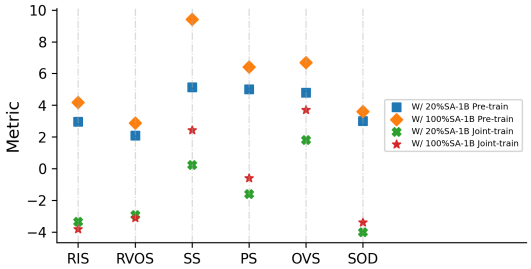


Figure 5. Effect of incorporating 20% as well as 100% SA-1B data into training process under pre-training and joint training strategy.

The vertical axis represents the performance increase or decrease compared to training without SA-1B data. It can be seen that the pre-training strategy performs significantly superior to joint training. We attribute this to the presence of heavy noise in the pseudo-labeled SA-1B data disrupting the normal training space. In addition, 100% SA-1B is significantly better than 20% under pre-training, but the phenomenon is not the same under joint training due to the larger noise distribution.

Effectiveness of multi-task joint training: We also test the performance gains resulting from multi-task joint training. From Figure 6 (a) we can see that vanilla multi-task joint training does not result in performance boost on all tasks without pseudo-labeled data aiding training. For short text caption tasks, multi-task joint training leads to decrease. We believe this is due to differences in visual and linguistic distribution across tasks. While with the large-scale pseudo-labeled data for pretraining, this problem can be greatly mitigated by superior initialization, as shown in Figure 6 (b).

4.5. Visualization Results

Figure 4 shows some segmentation examples for each task, which demonstrates the capability of our model to segment

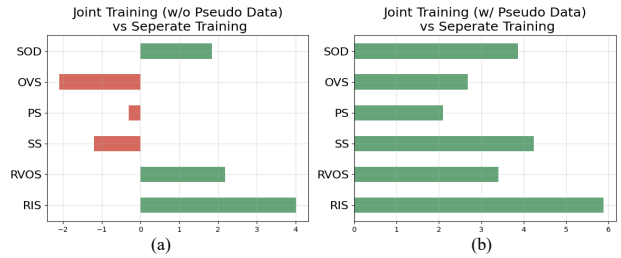


Figure 6. Influence of multi-task joint training. (a) shows the results without large-scale pseudo-labeled data for pre-training. (b) demonstrates results trained with pseudo-labeled data.

images at arbitrary semantic granularity with language instructions. Due to limited space, please see more results in supplementary materials.

5. Conclusion

In this paper we aim to achieve universal segmentation of arbitrary semantic level with language instruction. We reorganize a group of tasks from original diverse distributions into a unified data format for joint training, *i.e.*, triplet of images, masks, and captions. To promote the model's understanding of high-level language instructions, we present a fully aligned framework called UniLSeg. Combined with an automatic annotation engine for leveraging numerous unlabeled or weakly labeled data, our UniLSeg achieves superior performance on various semantic-related tasks.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62206153 and No. U1903213, in part by Shenzhen Science and Technology Program under Grant CJGJZD20220517142402006 and JCYJ20220818101014030.

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4975–4985, 2022. 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 4

[4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, pages 7454–7463, 2019. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[6] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, pages 17864–17875, 2021. 7

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 1, 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 3, 6

[10] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *CVPR*, pages 4964–4973, 2022. 6

[11] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 6

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 5, 6

[13] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, pages 15506–15515, 2021. 3, 6

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 4

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 3, 6

[16] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *ICCV*, pages 797–807, 2023. 1, 2, 3, 6

[17] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *ECCV*, pages 128–145. Springer, 2022. 2, 5, 7

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124, 2016. 3

[19] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020. 3

[20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020. 3

[21] Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. *arXiv preprint arXiv:2312.00869*, 2023. 2

[22] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75, 2020. 6

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 6

[24] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 3, 6

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 5, 6

[26] Yun Yi Ke and Takahiro Tsubono. Recursive contour-saliency blending network for accurate salient object detection. In *WACV*, pages 2940–2950, 2022. 1, 7

[27] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18145–18154, 2022. 3, 6

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[29] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019. 7

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 5, 7

[31] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3, 6

[32] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 5

[34] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 6

[35] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *TPAMI*, pages 871–885, 2018. 5

[36] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017. 7

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 5

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4

[39] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, pages 1280–1289, 2017. 3

[40] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 3, 5, 6

[41] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, pages 18653–18663, 2023. 3, 5, 6

[42] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 6

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[44] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, pages 648–665, 2022. 1

[45] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *ECCV*, pages 468–486, 2022. 1

[46] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. *arXiv preprint arXiv:2312.04089*, 2023. 1

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3, 5

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 2, 7

[49] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *ACM MM*, pages 1274–1282, 2020. 6

[50] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 6

[51] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *arXiv preprint arXiv:2305.17011*, 2023. 1, 2, 6

[52] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[53] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 5

[54] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 2, 5, 6

[55] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 7

[56] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 2

[57] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 2, 5

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 5, 6

[59] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18061–18070, 2022. 4

[60] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223, 2020. 2, 5, 6

[61] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, pages 629–645, 2020. 5

[62] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 5

[63] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, pages 38–54, 2018. 3

[64] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, pages 717–729, 2015. 5

[65] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 5

[66] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 2

[67] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2, 6, 7

[68] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 2, 3, 5, 6

[69] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 7

[70] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022. 6

[71] Qin Wu, Jianzhe Wang, Zhilei Chai, and Guodong Guo. Multi-scale feature aggregation and boundary awareness network for salient object detection. *Image and Vision Computing*, page 104442, 2022. 7

[72] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 7

[73] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 2, 3, 6

[74] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *CVPR*, pages 373–381, 2016. 1

[75] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, pages 15325–15336, 2023. 2, 5, 6

[76] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[77] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *CVPR*, pages 11266–11275, 2021. 3, 6

[78] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 1, 2, 3, 5, 6

[79] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 3

[80] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 3

[81] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 2, 5

[82] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020. 7

[83] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 5, 6

[84] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, pages 2337–2348, 2022. 4

[85] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 45(3):3738–3752, 2022. 1, 7

[86] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2, 5, 6